Statistiques de rangs

Y. BRUNET-MORET
Ingénieur hydrologue à l'O.R.S.T.O.M.

RÉSUMÉ

Cet article traite, sans épuiser le sujet, de quelques aspects de la statistique des rangs dans un échantillon tiré d'une distribution continue.

Chacun de ces aspects est examiné d'abord en supposant parfaitement connue la loi de répartition de la population mère, puis en supposant que les valeurs des paramètres d'une loi de répartition choisie ont été estimées d'après l'échantillon.

ABSTRACT

This paper deals, unthoroughly, with some aspects of order-statistics in a sample drawn from a continuous distribution.

Each aspect is examined at first supposing exactly known the distribution of the population, then supposing parameters values estimated by the sample.

SOMMAIRE

Cet exposé concerne plus spécialement les distributions de variables aléatoires absolument continues.

1. INTRODUCTION

Nous avons ici l'intention de faire le point sur différents problèmes qui sont liés par le fait qu'on y considère des échantillons de valeurs rangées, et que les rangs des valeurs y jouent souvent un rôle important.

Ces problèmes admettent des solutions différentes suivant que l'on suppose connue la distribution de la population mère (y compris les valeurs numériques des paramètres) ou qu'on en choisit la forme mathématique, les paramètres étant estimés d'après l'échantillon observé.

En premier lieu, nous proposons une solution à la question « quelle probabilité associer au rang d'ordre d'une valeur observée dans un échantillon de taille connue», question importante pour les représentations graphiques.

Puis nous étudions les distributions de quantiles échantillonnés, c'est-à-dire de valeurs correspondant à un rang donné dans des échantillons classés de taille n tirés au hasard dans une population mère parfaitement connue. Cette opération est analogue à l'étude de la distribution des intervalles de confiance de ces mêmes quantiles, avec les mêmes hypothèses.

Cependant, il vaut mieux à notre sens ne parler d'intervalles de confiance que lorsque la distribution de la population mère n'est pas connue a priori et qu'on a seulement choisi la forme mathématique de cette distribution. Nous proposons une méthode générale à base de tirages Monte Carlo.

Pour finir nous traitons, suivant les idées de M. BERNIER, du danger d'apparition et du temps de retour, en donnant à ce temps de retour une définition précise et restrictive, qui peut ne pas être acceptée pour le terme lui-même.

Ces exposés concernent plus spécialement les distributions de variables aléatoires absolument continues.

2. PROBABILITÉ DES VALEURS D'UN ÉCHANTILLON OBSERVÉ

Il est facile de représenter une loi de distribution, connue a priori ou ajustée à un échantillon, en portant les valeurs de la variate, calculées d'après la loi, en fonction de la fréquence au non dépassement (ou au dépassement). Il se pose un problème lorsqu'on veut représenter sur le même graphique les valeurs observées de l'échantillon.

Soit un échantillon de taille n de variables aléatoires indépendantes $x_1 \ldots x_i \ldots x_n$, rangées en ordre croissant par exemple; la question se pose de savoir quelle probabilité on va attribuer à chaque valeur de rang i pour reporter les points sur le graphique. Les opinions semblent assez divergentes suivant les auteurs. Nous proposons les solutions suivantes, en distinguant deux cas bien différents, et en supposant que la forme de la loi de distribution (connue ou choisie a priori) est une fonction continue de la variate, F(x) admettant une fonction de densité continue, f(x).

2.1. LOI DE DISTRIBUTION DE LA POPULATION MÈRE PARFAITEMENT CONNUE (par son expression mathématique *et* les valeurs numériques des paramètres)

Ce cas n'est pas celui qui nous intéresse le plus, mais nous l'exposons pour bien montrer la différence avec le cas qui correspond à l'ajustement des valeurs des paramètres d'après les valeurs de l'échantillon observé, sur une fonction de probabilité choisie a priori, avec ou sans justifications théoriques.

2.1.1. La probabilité pour que, dans un échantillon de taille n tiré de la population mère parfaitement connue, il y ait (i — 1) valeurs inférieures à une valeur x, qu'une valeur tombe dans l'intervalle $x \pm \frac{dx}{2}$, et que les (n — i) autres valeurs soient supérieures à x, est proportionnelle à :

$$[F(x)]^{i-1}\,[1 - F(x)]^{n-i}\,f(x)\,dx$$

et non égale à cette quantité du fait que l'amplitude de dx n'est pas définie.

On remarquera que la probabilité ainsi établie n'a de sens que si F(x) est complètement connue, de manière que, connaissant une valeur particulière x_i et son rang i, on puisse écrire cette expression de la probabilité sans avoir à se préoccuper des n — 1 autres valeurs de l'échantillon.

La loi de distribution de la valeur x_i de rang i dans les échantillons de taille n, loi définie par sa fonction de densité est donc :

$$dG_i = \frac{1}{B(i,\, n - i + 1)}\,[F(x_i)]^{i-1}\,[1 - F(x_i)]^{n-i}\,f(x_i)\,dx,$$

avec :

$$B(i, n - i + 1) = \frac{\Gamma(i)\ \Gamma(n - i + 1)}{\Gamma(n + 1)},$$

et la loi de distribution de la valeur F_i de la probabilité associée au rang i est :

$$dG_i = \frac{F_i^{i-1} (1 - F_i)^{n-i}\ dF}{B(i, n - i + 1)}$$

C'est une loi Bêta incomplète (puisque F suit une loi continue uniforme entre les valeurs 0 et 1) et la valeur moyenne de F_i est $\frac{i}{n + 1}$.

2.1.2. Cette valeur moyenne est souvent adoptée comme valeur de la fréquence associée à la variable de rang i pour report graphique. Il nous semble bien préférable de remplacer cette valeur moyenne par la valeur médiane F définie par :

$$\frac{1}{2} = \frac{1}{B(i, n + 1 - i)} \int_0^F F^{i-1} (1 - F)^{n-i}\ dF$$

qui est d'autant plus différente de la valeur moyenne que i est différent de $\frac{n}{2}$.

2.1.3. Cette valeur médiane est particulièrement facile à calculer pour le rang n ou le rang 1 :

$$F^n = \frac{1}{2} \qquad \text{et} \qquad F = \sqrt[n]{\frac{1}{2}}$$

d'où le tableau ci-dessous, où le temps de récurrence est défini par l'inverse de la fréquence au dépassement (si cette fréquence est inférieure à $\frac{1}{2}$).

Au rang n	Médiane Probabilité (au non dépassement)	$F = \left(\frac{1}{2}\right)^{1/n}$ Temps m de récurrence	Formule $\frac{n + 0.4}{i - 0,3}$ (i = 1)	Moyenne Probabilité (au non dépassement)	$F = \frac{n}{(n + 1)}$ Temps M de récurrence	Rapport $\frac{m}{M}$
Pour :						
n = 2	0,7071	3,42	3,43	0,6667	3	1,14
5	0,8705	7,73	7,72	0,8333	6	1,29
10	0,9330	14,9	14,85	0,9091	11	1,35
20	0,9659	29,4	29,2	0,9526	21	1,40
50	0,9862	72,5	72,0	0,9804	51	1,42
100	0,9931	144,9	143,5	0,9901	101	1,44
$n \to \infty$		$\to \frac{n}{\text{Log } 2}$	$\to \frac{n}{0,7}$		$\to n$	$\to 1,443$
		$\not\equiv \frac{n}{0,7}$	$\not\equiv \frac{n}{\text{Log } 2}\ (1 - 0,01)$			

On peut choisir, pour représenter la fréquence médiane associée au rang i, une formule approchée de la forme $\frac{i - a}{n + 1 - 2a}$ (respectant la symétrie par rapport à $\frac{n}{2}$) et en prenant a = 0,3 on est assuré de faire une erreur inférieure à 1%, quel que soit n, sur la récurrence correspondant au rang 1 ou au rang n. L'erreur est encore plus petite lorsque la valeur du rang i se rapproche de $\frac{n}{2}$.

2.1.4. D'où la règle : La fréquence au non dépassement associée à la valeur de rang i dans un échantillon rangé de taille n, tiré d'une population mère parfaitement connue, peut se représenter par :

$$\frac{i - 0{,}3}{n + 0{,}4}$$

2.2. LOI DE DISTRIBUTION ADAPTÉE A L'ÉCHANTILLON

La formulation mathématique de la loi de distribution est choisie a priori (ou connue à l'avance) et les valeurs numériques des paramètres (ou de certains paramètres seulement, mais au moins un) sont calculées d'après les valeurs de l'échantillon observé.

2.2.1. Nous faisons l'hypothèse suivante, probable et peut-être indémontrable : la méthode choisie pour le calcul des paramètres tend à rendre aussi voisine que possible d'une répartition continue et uniforme, la répartition des fréquences calculées. Le calcul de ces fréquences est fait, bien entendu, pour chaque valeur observée de l'échantillon, en appliquant la loi de forme choisie dont les paramètres ont été, cette fois, estimés à partir dudit échantillon, alors qu'en 2.1., leurs valeurs étaient fixées a priori.

On n'est pas ramené au problème précédent, car si la loi continue uniforme est parfaitement connue a priori, les fréquences calculées ne sont pas indépendantes des fluctuations de l'échantillonnage, ce qui les lie à l'estimation des paramètres, et empêche de pouvoir considérer une valeur individuelle de rang donné sans prendre en compte les n — 1 autres valeurs. La répartition de la fréquence associée à la valeur de rang i dans un échantillon rangé de taille n dépend donc certainement du nombre de paramètres calculés, probablement du mode de calcul des paramètres (*maximum de vraisemblance, moments...*) et peut-être de la *formulation mathématique* de la *fonction* de distribution choisie.

2.2.2. La dernière remarque ci-dessus sur la répartition des fréquences associées au rang i ne doit pas nous empêcher de tirer parti de l'hypothèse probable posée plus haut, et de chercher à calculer n fréquences obéissant le plus exactement possible à une répartition continue et uniforme.

On peut partir des considérations suivantes :

— d'une part, les fréquences $F_1 \ldots F_i \ldots F_n$ associées aux rangs 1 ... i ... n, sont comprises entre 0 et 1 et suivent une loi continue et uniforme. Or une telle loi peut être considérée comme une loi Bêta incomplète dégénérée de paramètre de position 0 et 1 et de paramètres de forme $p = 1$ et $q = 1$. Considérons maintenant que les F suivent effectivement une loi Bêta dont les paramètres de position sont connus a priori (0 et 1, intervalle de variation de F). Les équations du maximum de vraisemblance (voir [2] p. 63) :

$$\psi(p + q) - \psi(p) + \frac{1}{n} \sum \text{Log } F_i = 0$$

$$\psi(p + q) - \psi(q) + \frac{1}{n} \sum \text{Log } (1 - F_i) = 0$$

pour que l'on ait $p = 1$ et $q = 1$ (distribution uniforme), on doit donc avoir :

$$\sum_1^n \text{Log } F_i = -n \qquad \text{et} \qquad \sum_1^n \text{Log } (1 - F_i) = -n$$

équations en fait identiques à cause de la condition évidente de symétrie des F_i autour de la valeur $\frac{1}{2}$. Remarquons que les déterminations par le maximum de vraisemblance ne sont pas forcément absolument correctes, mais sûrement correctes ;

— d'autre part, le moment d'ordre r de la loi continue et uniforme est égal à $\frac{1}{r + 1}$ estimé par $\frac{1}{n} \sum_1^n F_i^r$.

Les fréquences F_i sont donc les racines de l'équation :

$$X^n + A_1 X^{n-1} + \ldots + A_{n-1} X + A_n = 0$$

les coefficients A étant déterminés par les moments d'ordre 1 à n (formule de Newton).

En effectuant le changement de variable :

$$x = 2F - 1$$

x suit une loi continue et uniforme dans l'intervalle $-1, +1$ et

$$S_{2r} = \sum_1^n x^{2r} = \frac{n}{2r+1}, \qquad \sum_1^n x^{2r+1} = 0$$

si n est pair, en posant $n = 2m$, les x sont racines de :

$$X^{2m} + A_2 X^{2m-2} + \ldots + A_{2p} X^{2m-2p} + \ldots + A_{2m-2} X^2 + A_{2m} = 0$$

si n est impair, en posant $n = 2m + 1$, il y a une racine $x = 0$ et les 2m autres racines sont celles de l'équation ci-dessus. Les coefficients A étant calculés par les formules de Newton :

$$S_{2p} + A_2 S_{2p-2} + \ldots + A_{2p}\, 2p = 0$$

Comme, après le changement de variable, $\Sigma \operatorname{Log} F_i = -n$ peut s'écrire $\Sigma \operatorname{Log}(1 + x_i) - n \operatorname{Log} 2 = -n$, où, en revenant aux valeurs naturelles :

$$\prod_1^n (1 + x_i) = 2^n e^{-n}$$

il vient, quelle que soit la parité de n :

$$\left(\frac{2}{e}\right)^n = 1 + A_2 + A_4 + \ldots + A_{2m-2} + A_{2m} \quad ([3], \text{p. } 339)$$

On peut, pour déterminer les coefficients A :

— soit utiliser le système linéaire formé par les m premières formules de Newton,

— soit utiliser le système linéaire formé par les $(m - 1)$ premières de ces formules et celle déduite de :

$$\sum_1^n \operatorname{Log} F_i = -n$$

Il semble bien que, si n est assez grand, ces deux systèmes soient équivalents. Il resterait à résoudre l'équation en X^{2m}, ce qui pourrait se faire par approximations successives en remarquant qu'on a une racine dans chaque intervalle $\frac{2(i-1)}{n} - 1, \frac{2i}{n} - 1$ (i de 1 à n).

2.2.3. Le procédé ci-dessus est bien compliqué, et nous allons chercher s'il est possible de trouver pour F_i une formule approchée de la forme $\frac{i-a}{n+1-2a}$, respectant la symétrie par rapport à $\frac{n}{2}$.

L'expression $\sum_1^n \operatorname{Log} F_i$ devient :

$$-n \operatorname{Log}(n + 1 - 2a) + \operatorname{Log} \Gamma(n + 1 - a) - \operatorname{Log} \Gamma(1 - a)$$

en développant $\operatorname{Log} \Gamma(n + 1 - a)$ suivant la formule de STIRLING, il vient, après avoir posé $z = n + 1 - a$:

$$-(z - 1 + a) \operatorname{Log}(z - a) - \operatorname{Log} \Gamma(1 - a) + \left(z - \frac{1}{2}\right) \operatorname{Log} z - n - 1 + a + \operatorname{Log} \sqrt{2\pi} + \frac{1}{12z} - \frac{1}{360z^3} \ldots$$

d'où :

$$\left(\frac{1}{2} - a\right) \operatorname{Log}(z - a) - \operatorname{Log} \Gamma(1 - a) - \left(z - \frac{1}{2}\right) \operatorname{Log}\left(1 - \frac{a}{z}\right) - n - 1 + a + \operatorname{Log} \sqrt{2\pi} + \frac{1}{12z} - \frac{1}{360z^3} \ldots$$

cette expression ne peut être voisine de $-n$ que si a est égal à $\frac{1}{2}$ et devient, après développement de $\operatorname{Log}\left(1 - \frac{1}{2z}\right)$:

$$-n + \operatorname{Log} \sqrt{2} - \frac{7}{24z} + \frac{5}{48z^2} \ldots \qquad \text{avec} \qquad \operatorname{Log} \sqrt{2} = 0.34657\ldots$$

La quantité $\text{Log}\sqrt{2} - \frac{7}{24\left(n-\frac{1}{2}\right)} + \frac{5}{48\left(n-\frac{1}{2}\right)^2}$, monotone et décroissante, est toujours beaucoup plus petite que les valeurs habituelles de n (environ 0,4 pour $n = 1$ et 0,35 pour $n \to \infty$).

D'autre part, on peut constater que le moment d'ordre 1, $\frac{1}{n}\sum_1^n \frac{i-\frac{1}{2}}{n}$, est égal à $\frac{1}{2}$ et que les moments d'ordre 2 et supérieurs $\frac{1}{n}\sum_1^n \left(\frac{i-\frac{1}{2}}{n}\right)^r$, tendent vers $\frac{1}{r+1}$ pour $n \to \infty$, les termes négligés dans ces moments étant d'ordre 2 et supérieurs en $\frac{1}{n}$. Les quantités $F_i = \frac{i-\frac{1}{2}}{n}$ sont donc des solutions approchées de l'équation en X^{2m} vue plus haut.

2.2.4. D'où la règle : La fréquence associée à la valeur de rang i dans un échantillon rangé de taille n, lorsque cet échantillon sert à calculer les valeurs des paramètres (et même d'un seul paramètre) d'ajustement d'une fonction de distribution, peut être estimée par $\frac{i-\frac{1}{2}}{n}$.

Il faut noter que les échantillons d'observations de données naturelles, comme les pluies ou les débits, relèvent toujours de ces hypothèses ; il est donc toujours nécessaire d'estimer les fréquences empiriques par $\left(\frac{i-\frac{1}{2}}{n}\right)$ *et non par* $\frac{i}{(n+1)}$ *ou une autre formule.*

2.3. EXTENSION

On peut se trouver devant le cas suivant, ou un cas analogue : nous devons, par exemple, faire l'étude des débits maximaux annuels à une station hydrométrique. Nous connaissons ces débits pour les $n = 20$ dernières années, observées, mais nous savons que pendant les $m = 50$ années précédant les observations, il y a eu $l = 2$ crues exceptionnelles (dans l années différentes), dont nous pouvons reconstituer les débits maximaux, que l'on sait être, pour chacun des $l = 2$, supérieurs à tous les $n + m - l$ autres débits maximaux annuels.

On choisit une forme mathématique de fonction de répartition. On ajuste ses paramètres à l'aide des n valeurs observées régulièrement dans la série complète. On trace ensuite la courbe représentative de cette loi de distribution (débits, fréquences), puis on porte sur le même graphique les n points observés, suivant la formule « fréquence expérimentale $= \frac{i-\frac{1}{2}}{n}$ ». Le problème est de représenter sur le graphique les l débits maximaux connus antérieurement à n, pour se rendre compte de leur position par rapport à la courbe.

Bien qu'il ne soit strictement pas possible de se considérer comme dans le cas traité en 2.1. (puisque n valeurs ont été utilisées pour déterminer les paramètres), on l'admettra comme hypothèse de travail et on affectera, aux l valeurs exceptionnelles, les fréquences qu'elles auraient eues si on avait tiré un échantillon de $(n + m)$ valeurs dans une population mère parfaitement connue ; on leur attribuera donc les fréquences (au dépassement) :

$$\frac{1-0,3}{n+m+0,4}, \frac{2-0,3}{n+m+0,4}, \ldots, \frac{l-0,3}{n+m+0,4}$$

Ce point de vue impose $\frac{l-0,3}{n+m+0,4} < \frac{1}{2n}$, et en fait que m soit du même ordre de grandeur que n, ou supérieur à n. Si m est inférieur à n, nous proposerions avec $l = 1$ d'utiliser la fréquence (au dépassement) de $\frac{1}{2(n+m)}$.

Il ne faut pas perdre de vue qu'il s'agit là d'un procédé empirique sans base théorique.

2.4. REPRÉSENTATION GRAPHIQUE EN TEMPS DE RÉCURRENCE

Dans la plupart des cas, une représentation graphique claire et parlante peut se faire sur papier semi-logarithmique, en utilisant l'échelle logarithmique comme support du « temps de récurrence ».

Ce temps de récurrence est défini, dans le cas de la représentation d'une fonction de distribution, comme l'inverse de la fréquence au non dépassement ou de la fréquence au dépassement lorsque cette fréquence est inférieure à $\frac{1}{2}$ (temps de récurrence de la médiane : 2 ans). Il évite d'utiliser systématiquement la fréquence au dépassement : la convention généralement admise veut que, lorsqu'on utilise le mot fréquence sans qualification, il s'agit de « fréquence au non dépassement ».

La figure 1 montre une telle représentation, sur papier semi-logarithmique, d'un échantillon de 28 valeurs auquel on a ajusté trois distributions différentes : le temps de récurrence pour les valeurs observées est calculé suivant la formule $\frac{n}{i - 0{,}5}$, i étant le rang du point à partir de la droite ou de la gauche, jusqu'à $i = \frac{n}{2}$.

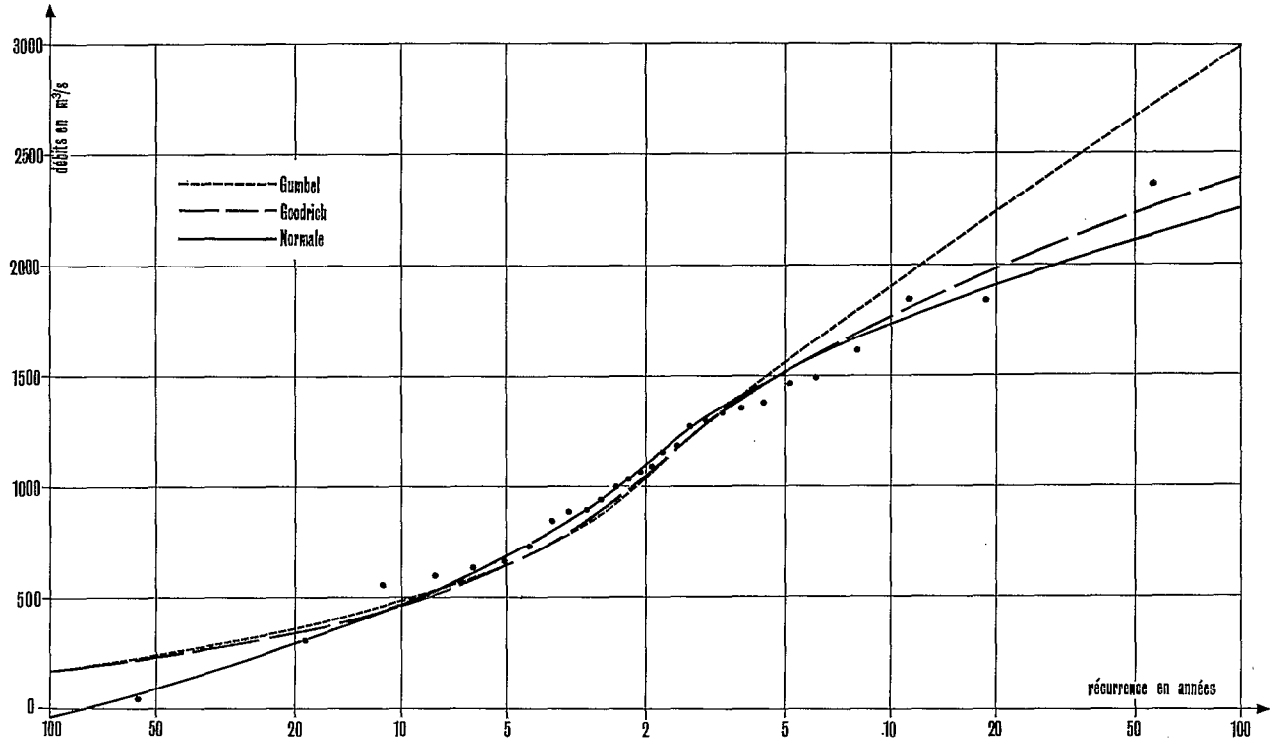

Fig. 1 — Oueme au pont de Savé échantillon de 28 débits maximaux annuels

3. DISTRIBUTION D'UN QUANTILE

Dans ce paragraphe, on traite uniquement le cas où la population mère est parfaitement connue, à la fois par la forme mathématique de la loi de distribution et par les valeurs numériques des paramètres de cette loi. Au paragraphe 4, on exposera le problème de la distribution des quantiles dans le cas où la forme de la loi est choisie d'avance, mais où les paramètres sont estimés à partir de l'échantillon : il s'agit alors de la notion d'intervalle de confiance.

Encore une fois, il est bien entendu que le premier cas n'est pas applicable aux observations de phénomènes naturels lets que pluies ou débits; l'exposé n'a pas d'autre but que celui de bien marquer la différence entre les deux phénomènes.

3.1. DISTRIBUTION D'UN QUANTILE. CAS GÉNÉRAL

3.1.1. Si on tire au hasard, indépendamment les unes des autres, n valeurs de la variable dans une population mère parfaitement connue (forme mathématique de la loi de distribution et valeurs numériques des paramètres connues a priori), les valeurs centrales et les quantiles calculés sur cet échantillon de taille n ne seront pas, sauf exceptions rares, égaux aux valeurs correspondantes, connues a priori, de la population mère.

Si nous tirons un nombre infini d'échantillons de même taille n dans cette population mère, et si nous rangeons chaque échantillon suivant les valeurs croissantes de la variable, les différentes valeurs de même rang i définissent une distribution (dépendant de la taille n). C'est ce genre de distribution, avec l'hypothèse préalable de population mère parfaitement définie, que nous nommons « distribution d'un quantile ». Le problème de l'« intervalle de confiance » en est très différent, nous le verrons au paragraphe 4.

3.1.2. Nous définissons le quantile comme étant la valeur de la variable de rang i dans un échantillon rangé de taille n (en lui associant la probabilité $\alpha = \frac{i}{n+1}$ au non dépassement, c'est-à-dire la probabilité $1 - \alpha = \frac{i'}{n+1}$ au dépassement, $i' = n + 1 - i$ étant le rang dans l'échantillon rangé en ordre décroissant). Si G(x), fonction continue et dérivable, représente la fonction de distribution de la population mère, la probabilité pour que, dans l'échantillon, il y ait (i — 1) valeurs inférieures à une valeur x, (n — i) valeurs supérieures à x, et une valeur comprise dans l'intervalle $x \pm \frac{dx}{2}$, est représentée par la probabilité élémentaire :

$$f(x)\,dx = \frac{1}{B(i,\, n+1-i)}\,[G(x)]^{i-1}\,[1 - G(x)]^{n-i}\,g(x)\,dx$$

Cette expression n'est pas souvent facile à intégrer pour obtenir la fonction de répartition du quantile.

3.1.3. On peut transcrire l'expression ci-dessus en :

$$dF = \frac{1}{B(i,\, n+1-i)}\,G^{i-1}\,(1 - G)^{n-i}\,dG$$

G étant une variate continue et uniforme dans l'intervalle 0,1.

La valeur moyenne de F est $\frac{i}{n+1}$ (ce qui ne veut pas dire que la valeur moyenne de x correspond à la fréquence $G(x) = \frac{i}{n+1}$).

La médiane de F est approximativement $\frac{i - 0{,}3}{n + 0{,}4}$ (la valeur médiane de x est donc, approximativement celle qui correspond à la fréquence $G(x) = \frac{i - 0{,}3}{n + 0{,}4}$).

Le mode de F est égal à $\frac{i-1}{n-1}$.

En prenant la dérivée logarithmique de f(x) par rapport à x, on trouve :

$$(i-1)\,\frac{g(x)}{G(x)} - (n-i)\,\frac{g(x)}{1 - G(x)} + \frac{g'(x)}{g(x)}$$

si $\frac{g'(x)}{g(x)}$ est négligeable par rapport à la somme des termes qui le précèdent (ce qui impose en général, dans le cas des distributions « en cloche », i assez voisin de $\frac{n}{2}$) le mode de la distribution du quantile est (approximativement) la solution de :

$$G(x) = \frac{i-1}{n-1}$$

3.1.4. On peut toujours obtenir la distribution du quantile par l'artifice suivant : On choisit une fréquence F, au non dépassement, qu'on porte dans la distribution du quantile. Par inversion de cette fréquence F dans la loi Bêta incomplète de paramètres de position zéro et un, et de paramètres de forme i et (n + 1 — i), on obtient une valeur G qui est une fréquence au non dépassement dans la loi connue de distribution de la population mère. Par inversion de cette fréquence G, on trouve la valeur de la variate qui correspond à la fréquence F choisie.

3.2. CAS PARTICULIER DE LA DISTRIBUTION NORMALE

Soit une population mère parfaitement connue suivant une distribution normale de moyenne x_0 et de variance σ^2 ; soit u_i la variable normale réduite correspondant à la fréquence $\frac{i}{n+1}$, la distribution du quantile de rang i dans des échantillons de taille n tirés de la population mère est celle de la quantité $y = \bar{x} + su_i$ où $\bar{x}$ et s sont deux variables aléatoires indépendantes (moyenne et écart-type d'un échantillon). Le cumulant d'ordre j de la distribution de y est égal au cumulant d'ordre j de la distribution de $\bar{x}$, plus u_i^j fois le cumulant d'ordre j de la distribution de s.

La variable $\bar{x}$ est distribuée suivant une loi normale de moyenne x_0 et de variance $\frac{\sigma^2}{n}$ dont tous les cumulants sont nuls à partir du troisième.

La variate s est distribuée suivant une loi de χ^2 à n degrés de liberté, de fonction de densité :

$$\frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{2}\chi^2\right)^{\frac{n}{2}-1}\exp\left(-\frac{1}{2}\chi^2\right)\mathrm{d}\left(\frac{\chi^2}{2}\right)$$

avec $\chi^2 = \frac{ns^2}{\sigma^2}$ dont les moments d'ordre j en s sont :

$$\left(\frac{2\sigma^2}{n}\right)^{j/2}\frac{\Gamma\left(\frac{n}{2}+\frac{j}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

en utilisant la formule de STIRLING pour développer l'expression de :

$$\frac{\Gamma\left(\frac{n}{2}+\frac{j}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

on obtient les expressions :

— de la moyenne : $\sigma\left(1-\frac{1}{4n}+\frac{1}{32n^2}+\frac{5}{128n^3}\cdots\right)$

— de la variance : $\frac{\sigma^2}{2n}\left(1-\frac{1}{4n}\cdots\right)$

— du cumulant d'ordre 3 : $\frac{\sigma^3}{4n^2}+\ldots$

— du cumulant d'ordre 4 : $\frac{3\sigma^4}{16n^4}+\ldots$

Les cumulants de la distribution de y sont donc, avec une bonne approximation :

— moyenne : $x_0 + u_i\,\sigma\left(1-\frac{1}{4n}\right)$

— variance : $\frac{\sigma^2}{n}\left(1+\frac{u_i^2}{2}\right)$

— cumulant d'ordre 3 : $\frac{\sigma^3}{4n^2}u_i^3$

— cumulant d'ordre 4 : $\frac{3\sigma^4}{16n^4}u_i^4$

d'où :

— coefficient d'asymétrie : $\frac{1}{\sqrt{2n}\left(\frac{2}{u_i^2}+1\right)^{3/2}}$

— coefficient d'aplatissement : $\dfrac{3}{4n^2\left(\dfrac{2}{u_i^2}+1\right)^2}$

ces coefficients ne sont jamais importants et la distribution de y peut être assimilée à une distribution normale.

Comme tous les cumulants de $\bar{x}$ et de s existent et remplissent les conditions nécessaires, on peut utiliser un développement en série limitée de fonction de distribution, pour avoir la distribution du quantile y. Si les fréquences les plus faibles et les plus fortes auxquelles on s'intéresse ne sont ni très petites ni très grandes, mettons comprises entre 0,05 et 0,95, il suffit probablement d'utiliser les 6 premiers cumulants des distributions de $\bar{x}$ et de s.

3.3. DISTRIBUTION ASYMPTOTIQUE DES VALEURS EXTRÊMES

En partant de (cf. 3.1.3.) :

$$\frac{1}{B(i, n+1-i)}\, G^{i-1}\,(1-G)^{n-i}\,dG$$

où G (x) est la fonction de distribution de la population mère, et en posant

si i = 1 $\quad y = nG \quad$ (valeur extrême inférieure)

si i = n $\quad y = n(1-G) \quad$ (valeur extrême supérieure)

nous obtenons :

$$dH(y) = n\left(1-\frac{y}{n}\right)^{n-1}\frac{dy}{n}$$

qui devient pour n grand :

$$dH(y) = e^{-y}\,dy$$

3.3.1. On démontre que, dans le cas de la valeur extrême supérieure, si $\dfrac{d}{dx}\left(\dfrac{1-G(x)}{G'(x)}\right)$ tend vers zéro lorsque x tend vers l'infini, hypothèse qui n'est pas vérifiée dans tous les cas, la distribution limite asymptotique de x_n, lorsque n tend vers l'infini, s'écrit :

$$F(x_n) = \exp\,[-\exp\,[-nG'(x_{0n})\,.\,(x_n - x_{0n})]]$$

où x_{0n} est défini par $G(x_{0n}) = 1 - \dfrac{1}{n}$.

De même, dans le cas de la valeur extrême inférieure, si $\dfrac{d}{dx}\left(\dfrac{G(x)}{G'(x)}\right)$ tend vers zéro lorsque x tend vers l'infini en valeurs négatives, la distribution limite de x_1 s'écrit :

$$F(x_1) = \exp\,[-\exp\,[-nG'(x_{01})\,.\,(x_{01} - x_1)]]$$

où x_{01} est défini par $G(x_{01}) = \dfrac{1}{n}$.

On trouve ainsi la distribution de GUMBEL, dont on notera le caractère arbitraire lorsqu'on l'utilise sans connaître la distribution G(x), ce qui est pratiquement toujours le cas dans le domaine qui nous occupe.

Le paramètre de position de cette distribution de GUMBEL est x_{0n} (ou x_{01}) et le paramètre d'échelle est $\dfrac{1}{[nG'(x_{0n})]}\left(\text{ou } \dfrac{1}{[nG'(x_{01})]}\right)$. Les valeurs numériques de ces paramètres sont donc connues a priori puisqu'on connaît parfaitement la distribution G(x), et le sont indépendamment de tout échantillon de taille n tiré de cette population mère.

Exemple : cas de la loi normale : l'expression $\dfrac{1-G(u)}{G'(u)}$ tend vers $\dfrac{1}{u}$ c'est-à-dire vers zéro, lorsque u tend vers l'infini, de même que sa dérivée qui tend vers $-\dfrac{1}{u^2}$.

La variable réduite tend vers $\sqrt{2\,\mathrm{Log}\,n}$ et x_{on} tend vers $x_0 + \sigma\sqrt{2\,\mathrm{Log}\,n}$, σ et x_0 étant les paramètres d'échelle et de position de la population mère. Comme $1 - G(u_{on})$ vaut $\dfrac{1}{n}$, $nG'(u_{on})$ tend vers u_{on} et $nG'(x_{0n})$ tend vers $\dfrac{\sqrt{2\,\mathrm{Log}\,n}}{\sigma}$.

Exemple : Dans la loi de GOODRICH, la fonction de répartition s'écrit $G(u) = 1 - e^{-u^{1/\delta}}$ avec $\delta > 0$, donc $\frac{1 - G(u)}{G'(u)}$ est égal à $\delta u^{1-\frac{1}{\delta}}$ dont la dérivée vaut $-\delta\left(1 - \frac{1}{\delta}\right)u^{-\frac{1}{\delta}}$; cette expression tend vers zéro lorsque u tend vers l'infini. Pour $G(u_{0n}) = 1 - \frac{1}{n}$, on a $u_{0n} = (\text{Log } n)^{\delta}$ et $nG'(u_{0n}) = \frac{1}{\delta}(\text{Log } n)^{1-\delta}$.

3.3.2. Si $\frac{d}{dx}\left(\frac{1 - G(x)}{G'(x)}\right)$ ne tend pas vers zéro lorsque x tend vers $+\infty$, dans le cas de la valeur extrême supérieure,

ou si $\frac{d}{dx}\left(\frac{G(x)}{G'(x)}\right)$ ne tend pas vers zéro lorsque x tend vers $-\infty$ dans le cas de la valeur extrême inférieure,

mais si, lorsque x tend vers $+\infty$, la valeur limite de l'expression $x^k(1 - G(x))$ est positive, finie et différente de zéro, c'est-à-dire si l'expression $\frac{xG'(x)}{1 - G(x)}$ tend vers une limite K positive,

ou si, lorsque $x \to -\infty$, la valeur limite de $|x|^k G(x)$ est positive, finie et différente de zéro, c'est-à-dire si l'expression $-\frac{xG'(x)}{G(x)}$ tend vers une limite k positive,

la distribution limite asymptotique lorsque n tend vers l'infini, s'écrit :

$$F(x_n) = \exp\left[-\left(\frac{x}{x_{0n}}\right)^{-k}\right]$$

ou :

$$F(x_1) = 1 - \exp\left[-\left(\frac{x}{x_{01}}\right)^{-k}\right]$$

avec $x_{0n} > 0$ défini par $G(x_{0n}) = 1 - \frac{1}{n}$

ou $x_{01} < 0$ défini par $G(x_{01}) = \frac{1}{n}$.

On reconnaît la distribution de FRÉCHET, cas particulier de la distribution exponentielle généralisée.

Exemple : Dans le cas de la loi de FRÉCHET $G(u) = e^{-u^{1/\delta}}$ avec $\delta < 0$. L'expression $u^k(1 - G(u))$ tend, lorsque u tend vers $+\infty$, vers $u^k u^{\frac{1}{\delta}}$ qui ne tend vers une limite positive finie et différente de zéro que pour $k = \frac{-1}{\delta}$; de même, la valeur limite de l'expression $\frac{uG'(u)}{[1 - G(u)]}$, lorsque u tend vers $+\infty$, est $k = \frac{1}{|\delta|}$: la distribution limite de la valeur extrême supérieure reste une loi de FRÉCHET de même paramètre de forme.

3.3.3. Si l'intervalle de définition de la population mère est borné par une valeur limite supérieure x_0, $G(x_0) = 1$ (qu'il y ait ou non une borne inférieure de l'intervalle de définition) et si les $(k - 1)$ premières dérivées de G(x) sont nulles pour $x = x_0$, la kième dérivée étant non nulle $G^{(k)}(x_0) \neq 0$ mais finie, et la $(k + 1)$ ième dérivée restant finie pour x différent de x_0, on peut développer $F(x_n) = \exp[-n(1 - G(x)]$ en série de TAYLOR à partir de x_0 et écrire la distribution limite, asymptotique pour n tendant vers l'infini, de la valeur extrême supérieure :

$$F(x_n) = \exp\left[-\left[\left(-\frac{n\,G^{(k)}(x_0)}{k!}\right)^{1/k}(x - x_0)\right]^k\right]$$

De même, si l'intervalle de définition de la population mère est borné par une valeur limite inférieure x_0 : $G(x_0) = 0$ (qu'il y ait ou non une borne supérieure de l'intervalle de définition) avec les mêmes conditions que ci-dessus pour les dérivées, on peut écrire la distribution limite asymptotique de la valeur extrême inférieure :

$$F(x_1) = \exp\left[-\left[\left(-\frac{n\,G^{(k)}(x_0)}{k!}\right)^{1/k}(x_0 - x)\right]^k\right]$$

Exemple : Dans le cas de la distribution exponentielle simple, $G(u) = 1 - e^{-u}$, la borne inférieure est $u_0 = 0$, la dérivée première de G(u) a une valeur finie non nulle pour $u = 0$: $G'(u_0) = 1$.

La distribution de la valeur extrême inférieure s'écrit :

$$F(u_1) = \exp\,[-\,[-\,n(u_0 - u)]] = 1 - \exp\,(-n\,u_1)$$

Dans le cas particulier, cette forme est valable quel que soit n, comme on peut le voir en partant de la définition :

$$dF(u_1) = \frac{1}{n}\,(1 - G)^{n-1}\,G'\,du$$

3.4. DISTRIBUTION CONJOINTE DE VARIABLES RANGÉES

Dans un échantillon rangé de n valeurs aléatoires et indépendantes tirées d'une population mère parfaitement connue par sa fonction de distribution G(x), la probabilité pour qu'il y ait (i — 1) valeurs inférieures à une valeur x_i, une valeur *comprise dans l'intervalle* $x_i \pm \frac{dx}{2}$, (j — i — 1) valeurs comprises entre x_i et x_j (j > i), une valeur comprise dans l'intervalle $x_j \pm \frac{dx}{2}$ et (n — j) valeurs supérieures à x_j s'écrit :

$$dF = \frac{G^{i-1}(x_i)\,[G(x_j) - G(x_i)]^{j-i-1}\,[1 - G(x_j)]^{n-j}\,dG(x_i)\,dG(x_j)}{B(i,\,j-i)\,B(j,\,n-j+1)}$$

3.4.1. Si nous définissons le pième quantile X_p de la population mère par $G(x_p) = P \quad (0 < P < 1)$ c'est-à-dire que 100 P % des valeurs de la population mère sont inférieures à x_p, cette quantité x_p ne peut se trouver dans l'intervalle x_i, x_j que si $G(x_i) \leqslant P \leqslant G(x_j)$ et la probabilité que cela se produise dans un échantillon de taille n est :

$$\sum_{k=i}^{j-1} \frac{n\,!}{k\,!\,(n-k)\,!}\,P^k\,(1 - P)^{n-k}$$

Si nous avons choisi des rangs symétriques j = n + 1 — i, la probabilité s'écrit :

$$\sum_{k=i}^{n-i} \frac{n\,!}{k\,!\,(n-k)\,!}\,P^k(1 - P)^{n-k}$$

Exemple : en prenant p = 0,5 et n pair, la probabilité pour que la médiane de la population mère se trouve dans l'intervalle $x_{n/2}$, $x_{n/2+1}$ est :

$$\frac{n\,!}{\frac{n}{2}\,!\,\frac{n}{2}\,!}\left(\frac{1}{2}\right)^n$$

pour n = 20 cette probabilité vaut 0,176. Lorsque n devient très grand, cette probabilité tend (en utilisant la formule de STIRLING) vers $\frac{1}{\sqrt{2\pi n}}$, et pour n = 2 000, la probabilité pour que la médiane de la population mère se trouve entre la 1 000e et la 1 001e valeurs rangées est d'environ 0,018.

3.4.2. Si dans l'expression :

$$dF = \frac{G(x_i)^{i-1}\,[G(x_j) - G(x_i)]^{j-i-1}\,[1 - G(x_j)]^{n-j}\,dG(x_i)\,dG(x_j)}{B(i,\,j-1)\,B(j,\,n-j+1)}$$

nous posons $y = n\,G(x_i)$

$z = n\,[1 - G(x_j)]$

il vient :

$$dF = \frac{\left(\frac{y}{n}\right)^{i-1}\left[1 - \frac{y}{n} - \frac{z}{n}\right]^{j-i-1}\left(\frac{z}{n}\right)^{n-j}\,dy\,dz}{B(i,\,j-1)\,B(j,\,n-j+1)\,n^2}$$

et si n tend vers l'infini, i et n — j restant fixes et petits, à la limite on peut écrire :

$$dF = \frac{y^{i-1}\, e^{-y}\, dy}{\Gamma(i)} \frac{z^{n-j}\, e^{-z}\, dz}{\Gamma(n - j + 1)}$$

c'est-à-dire qu'à la limite les distributions de y et de z sont indépendantes.

En faisant $i = 1$ et $j = n$, les cumulants d'ordre k des distributions de $y = n\, G(x_1)$ et de $z = n - n\, G(x_n)$ sont, à la limite lorsque n tend vers l'infini, égaux à $(k - 1)$!. Dans les mêmes conditions, le cumulant d'ordre k de la distribution de :

$$V = \frac{z + y}{n} = \frac{1}{n} [n - n\, [G(x_n) - G(x_1)]]$$

est égal à $(k - 1)\, ! \frac{2}{n^k}$, la quantité $(1 - V)$ représente l'étendue de l'échantillon exprimée en fréquence, calculée par la fonction de distribution de la population mère. V est distribuée suivant une loi gamma incomplète de paramètre de forme 2, de paramètre d'échelle n et de paramètre de position zéro ; son mode est égal à $\frac{1}{n}$, sa moyenne à $\frac{2}{n}$ et sa variance à $\frac{2}{n^2}$ (pour n grand).

La distribution de l'étendue proprement dite de l'échantillon $R = x_n - x_1$ n'est pas facile à étudier dans le cas général. Des tables ont été établies pour la distribution normale.

La fonction de distribution de la population mère étant G(x), celle de l'étendue R est, quel que soit n :

$$F(R) = n \int_{-\infty}^{+\infty} [G(x + R) - G(x)]^{n-1}\, dG(x)$$

4. INTERVALLE DE CONFIANCE

Le problème se pose ainsi dans sa généralité :

Nous avons un échantillon de taille n de valeurs observées, nous avons choisi a priori (ou nous connaissons à l'avance) la forme mathématique de la loi de distribution de la population mère dont provient l'échantillon, et nous avons calculé les valeurs numériques des paramètres inclus dans cette loi grâce aux valeurs observées de l'échantillon.

Quel intervalle de confiance peut-on associer à une valeur déduite mathématiquement de la distribution ainsi ajustée ?

4.1. DÉFINITION DE LA NOTION D'INTERVALLE DE CONFIANCE

4.1.1. Si nous pouvions faire un grand nombre de tirages d'échantillons de taille n dans la population mère, nous pourrions ajuster les valeurs numériques des paramètres de la forme mathématique de la loi de distribution choisie ou connue (nous pourrions connaître les distributions conjointes des valeurs numériques des paramètres) et pour chaque ajustement calculer la même valeur, par exemple un débit de crue de fréquence centenaire. L'intervalle de confiance à tant pour cent de la crue centenaire est défini par les valeurs limites inférieure et supérieure, entre lesquelles se trouvent tant pour cent des valeurs calculées du débit de crue centenaire.

On prend, d'habitude mais rien n'y oblige, un intervalle de confiance symétrique : l'intervalle de confiance à P pour cent du débit de crue centenaire a pour valeur limite inférieure celle qui correspond à la probabilité au non dépassement de $\frac{1}{2} - \frac{P}{2}$, et pour valeur limite supérieure celle qui correspond à la probabilité au non dépassement de $\frac{1}{2} + \frac{P}{2}$ (ou à la probabilité au dépassement de $\frac{1}{2} - \frac{P}{2}$).

On comprend que, pour une taille donnée d'échantillon et pour une fonction de répartition choisie, l'intervalle de confiance dépende de la façon de calculer les valeurs numériques des paramètres.

4.1.2. Il est quasiment impossible de partir des distributions conjointes des paramètres (distributions déterminées d'une part par la forme mathématique de la fonction de répartition choisie pour représenter la population mère, d'autre part par les valeurs numériques des paramètres ajustés à l'échantillon et la méthode d'ajustement) pour

déterminer les intervalles de confiance, et nous proposons pour ce faire la méthode générale suivante, valable pour n'importe quelle valeur déduite mathématiquement de la distribution ajustée. Soit, par exemple, à déterminer l'intervalle de confiance à 90%, symétrique, d'un débit de crue centenaire, d'après un échantillon observé de n valeurs aléatoires et indépendantes de débits maximaux annuels de crues, en admettant une distribution de GOODRICH. On calcule les valeurs numériques des paramètres d'ajustement de la loi à l'échantillon par la méthode du maximum de vraisemblance. Puis, dans la population mère de répartition définie par la loi de GOODRICH et ces valeurs numériques des paramètres, nous tirons un certain nombre d'échantillons, mettons 200, chacun de n valeurs aléatoires et indépendantes. Sur chacun de ces échantillons, nous calculons de nouvelles valeurs numériques des paramètres d'ajustement à une loi de GOODRICH (par la méthode du maximum de vraisemblance), puis la valeur du débit de crue centenaire au moyen de cette loi avec ces nouvelles valeurs des paramètres. Après avoir rangé en ordre croissant les 200 valeurs ainsi obtenues, nous donnerons l'intervalle de confiance à 90% comme étant compris entre la dixième et la cent quatre-vingt-dixième valeur.

4.2. CAS PARTICULIER DE LA DISTRIBUTION NORMALE

Nous avons choisi une loi normale pour représenter la population mère. Sur l'échantillon de taille n, nous avons calculé la moyenne :

$$\bar{x} = \frac{1}{n} \sum x_j$$

et la variance :

$$s^2 = \frac{1}{n-1} \sum (x_j - \bar{x})^2$$

La probabilité pour qu'une valeur quelconque x_0 soit la véritable moyenne de la population mère (étant donné que nous avons choisi une loi normale) est donnée par une loi de STUDENT, de densité de probabilité (en posant $t = \frac{x_0 - \bar{x}}{\sqrt{s^2/n}}$) :

$$\frac{\Gamma(n/2)}{\sqrt{n-1}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} \frac{dt}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}}$$

Les moments pairs d'ordre j ($j < n - 1$) ont pour expression :

$$\frac{(n-1)^{j/2}}{\Gamma\left(\frac{1}{2}\right)} \frac{\Gamma\left(\frac{j+1}{2}\right)\Gamma\left(\frac{n-1-j}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

dont on déduit les cumulants de la distribution de x_0 :

— moyenne : $\bar{x}$

— variance : $\frac{s^2}{n} \frac{n-1}{n-3}$

— cumulant d'ordre 3 : zéro

— cumulant d'ordre 4 : $\frac{6s^4(n-1)^2}{n^2(n-3)^2(n-5)}$.

La probabilité pour que σ^2 soit la véritable variance de la population mère (étant donné que nous avons choisi une loi normale) est représentée par une loi de χ^2 à $(n - 1)$ degrés de liberté, de densité de probabilité (en posant $\chi^2 = \frac{(n-1)\,s^2}{\sigma^2}$) :

$$\frac{1}{\Gamma\left(\frac{n}{2} - \frac{1}{2}\right)} \left(\frac{1}{2}\chi^2\right)^{\frac{n}{2} - \frac{3}{2}} \exp\left(-\frac{1}{2}\chi^2\right) d\left(\frac{1}{2}\chi^2\right)$$

dont les moments d'ordre j en σ sont, pour $j < (n - 1)$:

$$\left[\frac{(n-1)\,s^2}{2}\right]^{j/2} \frac{\Gamma\left(\frac{n}{2} - \frac{1}{2} - \frac{j}{2}\right)}{\Gamma\left(\frac{n}{2} - \frac{1}{2}\right)}$$

dont on déduit les cumulants de la distribution de σ :

— moyenne : $\frac{n-1}{n-2}\, s\left(1 - \frac{1}{4(n-1)} \cdots\right)$

— variance : $\frac{(n-1)^2}{2(n-2)^2\,(n-3)}\, s^2(1 - \ldots)$

— cumulant d'ordre 3 : $\frac{(n-1)^2}{4(n-2)^3\,(n-3)\,(n-4)}\, s^3\left(5n - \frac{17}{4}\right)(1 + \ldots)$

— cumulant d'ordre 4 : $\frac{(n-1)^4}{32(n-2)^4\,(n-3)^2\,(n-4)\,(n-5)}\, s^4\left(47n - \frac{1\,135}{8}\right)(1 + \ldots)$

u_i étant la valeur de la variable normale réduite correspondant à une probabilité P_i, l'estimation de la variable, non réduite, de même probabilité est $\bar{x} + su_i$, $\bar{x}$ et s ayant été déterminés par l'échantillon. L'intervalle de confiance est donné par la distribution de la variable aléatoire $v = x_0 + \sigma u_i$.

Comme x_0 et σ sont des variables aléatoires indépendantes, les cumulants d'ordre j de la distribution de v sont égaux à la somme des cumulants de mêmes ordres de la distribution de x_0, plus u_i^j fois les cumulants de mêmes ordres de la distribution de σ, d'où :

— moyenne de v : $\bar{x} + \frac{n-1}{n-2}\, su_i$

et pour n grand $\rightarrow \bar{x} + su_i$

— variance : $\frac{s^2(n-1)}{n-3}\left[\frac{1}{n} + u_i^2\,\frac{(n-1)}{2(n-2)^2}\right]$

et pour n grand : $\rightarrow \frac{s^2}{n}\left[1 + \frac{u_i^2}{2}\right]$

— cumulant d'ordre 3 : $\not\rightarrow s^3\,\frac{5(n-1)^3\,u_i^3}{4(n-2)^3\,(n-3)\,(n-4)}$

et pour n grand : $\rightarrow \frac{5s^3}{4n^2}\,u_i^3$

— cumulant d'ordre 4 : $\not\rightarrow \frac{s^4(n-1)^2}{(n-3)\,(n-5)}\left[\frac{6}{n^2(n-3)} + \frac{47u_i^4(n-1)^2}{32(n-2)^4\,(n-4)}\right]$

et pour n grand : $\rightarrow \frac{s^4}{n^3}\left[6 + \frac{47}{32}\,u_i^4\right]$

— coefficient d'asymétrie pour n grand : $\rightarrow \frac{5}{4\sqrt{n}}\,\frac{1}{\left(1 + \frac{1}{2}\,u_i^2\right)^{\frac{3}{2}}}$ avec le signe de u_i ;

— coefficient d'aplatissement pour n grand : $\rightarrow \frac{1}{n}\,\frac{\frac{6}{u_i^4} + \frac{47}{32}}{\left(1 + \frac{1}{2}\,u_i^2\right)^2}$.

Ces coefficients d'asymétrie et d'aplatissement ne sont jamais très différents de zéro, même lorsque n n'est pas très grand, et bien que tous les cumulants de v n'existent pas (ils disparaissent à partir de l'ordre n — 1), on peut utiliser un développement en série limitée de fonction de distribution en utilisant les 6 premiers cumulants de x_0 et de σ, si l'intervalle de confiance à calculer n'est pas trop grand, mettons inférieur ou égal à 90%. On peut plus simplement considérer que la distribution de v est normale.

4.3. CAS DE LA DISTRIBUTION DE GUMBEL

4.3.1. Si nous utilisons une distribution dont on ne détermine, à partir d'un échantillon observé, que les paramètres « x_0» de position et « s» d'échelle (les paramètres de forme n'existant pas ou leurs valeurs numériques étant connues ou fixées à l'avance) nous pouvons espérer trouver une méthode analogue à celle que nous venons de voir

dans le cas de la distribution normale : si u_i est la variable réduite correspondant à la probabilité P_i, la variable non réduite a pour valeur $x_i = x_0 + su_i$.

Si x_0 et s sont des variables indépendantes dont on peut connaître la distribution, le cumulant d'ordre j de la distribution de x_i sera la somme du cumulant de même ordre de la distribution de x_0 plus u_i^j fois le cumulant de même ordre de la distribution de s.

Il n'y a que dans le cas de la distribution normale que ces distributions de x_0 et de s sont indépendantes quel que soit la taille n de l'échantillon. Mais nous avons vu, dans « l'estimation des paramètres» (paragraphes 6.4. et 7.5.), que pour n grand on peut rendre nulle la covariance de x_0 et s (aux termes d'ordres $\left(\frac{1}{n}\right)^2$ et supérieurs près) en faisant un changement d'origine sur la variable réduite $\frac{x_i - x_0}{s} - u_0$.

4.3.2. Examinons le cas de la distribution de GUMBEL, dont la variable réduite u a pour cumulants :

— moyenne : $k_1 = 0{,}577\,2157 = c$ (constante d'EULER) ;
— variance : $k_2 = 1{,}644\,934$;
— cumulant : $k_3 = 2{,}404\,116$,
$k_4 = 6{,}493\,939$.

Il ne semble pas possible (c'est théoriquement faisable) de trouver les distributions des paramètres x_0 et s si on les détermine par la méthode du maximum de vraisemblance. En utilisant la méthode des moments (ou des cumulants) on obtient, en posant R_1 moyenne de l'échantillon, R_2 variance de l'échantillon :

$$s = \sqrt{\frac{R_2}{K_2}}$$

$$x_0 = R_1 - K_1\sqrt{\frac{R_2}{K_2}}$$

et pour n grand :

$$\text{— var (s)} \qquad = \frac{s^2}{4nK_2^2}\left[K_4 + 2\frac{n}{n-1}K_2^2\right] = \frac{s^2}{4nK_2^2}A$$

$$\text{— var } (x_0) \qquad = \frac{s^2}{n}\left[K_2 - \frac{K_3K_1}{K_1} + \frac{K_1^2}{4K_2^2}\left(K_4 + 2\frac{n}{n-1}K_2^2\right)\right] = \frac{s^2}{n}B$$

$$\text{— covar } (x_0, s) = \frac{s^2}{4nK_2}\left[2K_3 - \frac{K_1}{K_2}\left(K_4 + 2\frac{n}{n-1}K_2^2\right)\right] = \frac{s^2}{4nK_2}C$$

Le coefficient de corrélation linéaire entre x_0 et s s'écrit $\frac{C}{2\sqrt{AB}}$.

— Pour $n > 4$, le terme C est positif et croissant vers la valeur asymptotique de 0,63055 lorsque n croît.

— Les termes A et B sont positifs et décroissants vers les valeurs asymptotiques respectives de 11,905 et de 1,1678 lorsque n croît.

— La valeur maximale du coefficient de corrélation est obtenue asymptotiquement et elle est de 0,0845. Cette valeur est très petite bien qu'approximative pour n fini. Nous allons négliger la corrélation entre x_0 et s et considérer que les estimations de ces paramètres sont indépendantes.

La fonction de densité de la loi de GUMBEL s'écrit :

$$e^{-\frac{x-x_0}{s}} \exp\left(-e^{-\frac{x-x_0}{s}}\right)\frac{d(x-x_0)}{s}$$

et l'estimation de x_0 est $R_1 - K_1 s$. Comme x_0 et s sont considérés comme indépendants, on peut obtenir la distribution de $\frac{x_0}{s}$ (en considérant que s est une constante) et le premier cumulant de $\frac{x_0}{s}$ a pour valeur $\frac{R_1}{s} - c$ et les autres ont les mêmes valeurs que ceux de $\frac{x}{s}$.

$\bar{x}$ étant la moyenne de n variables aléatoires distribuées suivant une loi de GUMBEL, le cumulant d'ordre j de $\frac{\bar{x}}{s}$ est $\left(\frac{1}{n}\right)^{j-1}$ fois le cumulant d'ordre j de la distribution de la variate réduite, c'est-à-dire $\frac{(j-1)!}{n^{j-1}}\zeta(j)$, $\zeta(j)$ étant la fonction de RIEMANN. Les cumulants de la distribution de x_0 sont donc :

— le premier : $R_1 - c\ s$

— les suivants : $s^j \frac{(j-1)!}{n^{j-1}}\zeta(j)$.

L'estimation de s est $\sqrt{\frac{R_2}{\zeta(2)}}$. Il nous faudrait donc chercher la distribution de $\sqrt{R_2}$, en passant d'abord par celle de $v = \left(\frac{x - x_0}{s}\right)^2 = u^2$, s étant considéré comme une constante. Ayant les cumulants de la distribution de v, on pourrait avoir ceux de la distribution de $w = \sum_1^n u_i^2$, puis ceux de $R_2 = \frac{s^2}{n-1}(w - nc^2)$. De la distribution de R_2 on pourrait déduire les cumulants de la distribution de $\frac{\sqrt{R_2}}{\sqrt{\zeta(2)}}$. Il semble suffisant de s'en tenir aux quatre premiers cumulants de la variate $s = \frac{\sqrt{R_2}}{\sqrt{\zeta(2)}}$ (car n est supposé grand), et de les utiliser avec les cumulants de la variate x_0 dans un développement en série limitée de la fonction de distribution de $x_i = x_0 + s\ u_i$.

5. DANGER D'APPARITION

Nous supposons parfaitement connue la population mère par la forme mathématique de la fonction de probabilité au non-dépassement G(x) admettant une fonction de densité continue g(x), et par les valeurs numériques de ses paramètres. A une certaine fréquence choisie a priori, correspond dans cette loi une valeur parfaitement définie de la variable.

On veut savoir quelle peut être la fréquence d'apparition de cette valeur dans un nombre d'observations donné a priori : par exemple, fréquence d'apparition d'une crue millénaire en 100 années consécutives.

Nous avons déjà vu ce problème, sous une forme à peine différente, dans le paragraphe 3.

5.1. FRÉQUENCE D'APPARITION D'UNE VALEUR CONNUE A PRIORI

Dans l'échantillon de taille n, taille donnée a priori, la probabilité de n'avoir aucune observation égale ou supérieure à une valeur x définie par $G(x_p) = 1 - P$, $0 < P < 1$ étant la probabilité au dépassement égale à $\frac{1}{T}$ où $T > n$ est le « temps de récurrence», est :

$$dF(x_p) = \frac{1}{n}[G(x_p)]^{n-1}\, dG(x_p)$$

$$F(x_p) = [G(x_p)]^n$$

et la probabilité d'avoir une ou plusieurs observations égales ou supérieures à x_p est :

$$1 - [G(x_p)]^n = 1 - \left(1 - \frac{1}{T}\right)^n$$

Pour $T = 1\,000$ et $n = 100$, $F(x_p) = 0{,}905$, c'est-à-dire que la probabilité de trouver en 100 années consécutives, une crue maximale annuelle de fréquence millénaire ou plus rare est de 9,5 %.

Le raisonnement est le même lorsqu'on traite une valeur x_p définie par une faible probabilité P au non dépassement $\left(\text{temps de récurrence } T = \frac{1}{P}\right)$. La probabilité d'avoir une ou plusieurs observations égales ou inférieures à x_p dans un échantillon de taille n est alors :

$$F(x_p) = 1 - \left(1 - \frac{1}{T}\right)^n$$

Par exemple, pour T = 100 et n = 20, on a $F(x_p) = 0{,}182$, c'est-à-dire que la probabilité de trouver en 20 années successives un total déficitaire annuel de précipitations de fréquence centenaire ou plus rare est de 18,2%.

5.2. PROBLÈME INVERSE

Inversement, nous pouvons calculer la valeur de la variable correspondant à une probabilité d'apparition A choisie à l'avance. Cette valeur est définie par :

$$G(x_p) = A^{1/n}$$

pour les valeurs élevées de la variable.

Ainsi, pour $n = 100$, $A = \frac{1}{2}$, il vient $G(x_p) = 0{,}9931$ c'est-à-dire que nous avons une probabilité de 0,5 de ne pas observer en 100 ans consécutifs une crue maximale annuelle, de fréquence supérieure ou égale à 0,9931 (ce qui correspond à un temps de récurrence de 145 ans).

Pour n = 100, A = 0,9, il vient $G(x_p) = 0{,}99895$, c'est-à-dire que nous avons une probabilité de 90% de ne pas observer en 100 ans consécutifs de crue maximale annuelle de fréquence supérieure ou égale à 0,99895 (ce qui correspond à un temps de récurrence de 952 ans).

Si on s'intéresse aux faibles valeurs de la variable, elle sera définie par :

$$G(x_p) = 1 - (1 - A)^{1/n}$$

Ainsi pour n = 20, A = 0,2, il vient $G(x_p) = 0{,}0111$, c'est-à-dire qu'il y a une probabilité de 20% de ne pas observer en 20 années consécutives un total annuel déficitaire de précipitations, de fréquence inférieure ou égale à 0,0111 — donc de temps de récurrence supérieur ou égal à 90 ans.

Nota : Evidemment, dans ce qui précède, nous supposons que les crues maximales annuelles ou les totaux annuels de précipitations sont des variables aléatoires et indépendantes.

6. TEMPS DE RETOUR

Il nous semble nécessaire de faire une distinction très nette entre le « temps de récurrence» et le « temps de retour».

Le temps de récurrence est défini d'après la loi de répartition de la population mère : inverse de la fréquence au non dépassement si celle-ci est inférieure à $\frac{1}{2}$, inverse de la fréquence au dépassement si celle-ci est inférieure à $\frac{1}{2}$.

Le temps de récurrence d'une valeur donnée x_p (supérieure à la médiane) est le nombre moyen d'unités de temps (c'est-à-dire le nombre moyen d'observations aléatoires et indépendantes n_i) que l'on trouve entre deux observations supérieures ou égales à x_p (une de ces observations étant comptée dans n_i). Autrement dit, dans un grand nombre de siècles, on observera en moyenne tous les cent ans un total pluviométrique annuel supérieur ou égal au total annuel excédentaire centenaire.

6.1. DÉFINITION

Le danger d'apparition est calculé d'après le temps de récurrence dans un laps de temps sans référence à la série chronologique observée : il ne dépend que de la fonction de répartition choisie et ajustée.

Nous définissons le temps de retour par rapport à cette série chronologique observée : dans combien de temps (ou à combien d'observations supplémentaires, que l'on ne possède pas encore) risque-t-on de trouver une valeur supérieure ou égale (pour les fortes valeurs à faible fréquence au dépassement) inférieure ou égale (pour les faibles valeurs à faible fréquence au non dépassement) à une des valeurs observées dans la série que l'on possède : en fait, la plus forte (ou la plus faible).

Pour être clair, nous allons considérer le cas concret d'une série chronologique de crues maximales annuelles, dans laquelle nous avons observé une crue x de récurrence très rare T, dont la fréquence au dépassement est très petite $[1 - G(x_p)]$ ($G(x_p)$ étant la fonction de distribution de la population mère) de façon qu'on puisse assimiler la variable t, temps séparant deux crues de fréquence supérieure ou égale à $G(x_p)$, à une variate exponentielle :

$$\text{Prob}\,(X \geqslant x_p) = [1 - (1 - G(x))]^t \to e^{-\frac{t}{T}} \text{ (au dépassement)}$$

(la valeur moyenne de la variable t est le temps de récurrence).

On démontre (étude probabiliste de processus stochastiques de renouvellement) que la probabilité de revoir une crue supérieure ou égale (en maximal annuel) à la crue x_p est celle de la variable θ (temps compté depuis l'apparition de la crue x_p), probabilité définie pour T et θ grands par :

$$F(\theta) = \int_0^\theta [1 - G(x_p)]t\, e^{-t(1-G(x_p))}\, dt = \int_0^\theta \frac{t}{T} e^{-\frac{t}{T}}\, dt$$

dont la valeur moyenne est 2T.

6.2. APPLICATION

On peut définir la probabilité expérimentale au dépassement de la valeur x_p, valeur maximale d'un échantillon chronologique de taille n, par :

$$\text{Prob}\,(X \geqslant x_p) = \frac{1}{n}$$

et le temps de récurrence de cette valeur x_p sera, en unités de temps, égal à n.

Si n est grand, la valeur moyenne du temps de retour de la valeur x_p sera égale à 2n, résultat, que l'on peut rapprocher de celui du paragraphe 2.2.4.

BIBLIOGRAPHIE

[1] BERNIER (J.) – Sur quelques difficultés rencontrées en hydrologie statistique dans le calcul d'un quantile. E.D.F. Chatou.

[2] BERNIER (J.) – 1967 – Sur la théorie de renouvellement et son application en hydrologie. E.D.F. Chatou.

[3] BRUNET-MORET (Y.) – 1969 – Etude de quelques lois statistiques utilisées en hydrologie. *Cah. O.R.S.T.O.M., sér. Hydrol.*, vol. VI, n° 3-69.

[4] KENDALL (M.G.) and STUART (A.) – 1963 – The advanced theory of statistics. Vol. 1, 2e édition, éd. Griffin, Londres.

En particulier : § 3.1., cf. chapitre 10 (KENDALL), §3.3., cf. chapitre 14 (KENDALL), § 3.4., cf. chapitres 14 et 32 (KENDALL).

[5] WARUSFEL (A.) – 1966 – Dictionnaire raisonné de mathématiques. Ed. du Seuil, Paris.